NORTHWESTERN POLYTECHNIC UNIVERSITY

Overfitting To Evaluate Linear and NonLinear Regression Models

Prepared For:

Mr. Henry Chang
Machine Learning and Business Intelligence CS550
Summer 2021
Northwestern Polytechnic University

Prepared By:

Ms. Nagalla, Santhi Sree ID:19568

Table of Contents

Sample Data Distribution

The concept of Training/Cross-Validation/Test Data Sets is when you have a large data set, it's recommended to split it into 3 parts:

- Training set (50% of the original data set): This is used to build up our prediction algorithm. Our algorithm tries to tune itself to the quirks of the training data sets. In this phase we usually create multiple algorithms in order to compare their performances during the Cross-Validation Phase.
- Cross-Validation set (25% of the original data set): This data set is used to compare the performances of the prediction algorithms that were created based on the training set. We choose the algorithm that has the best performance.
- Test set (25% of the original data set): Now we have chosen our preferred prediction algorithm but we don't know yet how it's going to perform on completely unseen real-world data. So, we apply our chosen prediction algorithm on our test set in order to see how it's going to perform so we can have an idea about our algorithm's performance on unseen data.

Three Phases of Sample Data

Training Phase				Validation Phase				Test Phase	
Real Data Set 1 50% of the collcted data		Model 1: Linear Regression	Model 2: Non-Linear Regression	Real Data Set 2 25% of the collcted data		Model 1: Linear Regression	Model 2: Non-Linear Regression	Real Data Set 3 25% of the collcted data	The better model (Model 1 or Model 2) selected from the Validation Phase based on the analysis of overfitting will be used to calculate $\hat{y}$
	▪ After calculating a1, b1, a2, b2 in Training Phase, the values are not changed with the new Real Data Sets in Validation Phase and Test Phase. ▪ Only $\hat{y}$ values are changed with the new Real Data Sets.								
x	y	$\hat{y}=a1 + b1 * x$	$\hat{y}=a2 + b2 * x^2$	x	y	$\hat{y}=a1 + b1 * x$	$\hat{y}=a2 + b2 * x^2$	x	$\hat{y}=a1 + b1 * x$ or $\hat{y}=a2 + b2 * x^2$
1	1.8			1.5	1.7			1.4	
2	2.4			2.9	2.7			2.5	
3.3	2.3			3.7	2.5			3.6	
4.3	3.8			4.7	2.8			4.5	
5.3	5.3			5.1	5.5			5.4	
1.4	1.5			X	X	X	X	X	X
2.5	2.2			X	X	X	X	X	X
2.8	3.8			X	X	X	X	X	X
4.1	4.0			X	X	X	X	X	X
5.1	5.4			X	X	X	X	X	X

Linear Regression Model

- Linear regression models are used to show or predict the relationship between two variables or factors. The factor that is being predicted (the factor that the equation solves for) is called the dependent variable. The factors that are used to predict the value of the dependent variable are called the independent variables.

Simple Linear Regression

- In linear regression, each observation consists of two values. One value is for the dependent variable and one value is for the independent variable. In this simple model, a straight line approximates the relationship between the dependent variable and the independent variable.
- When two or more independent variables are used in regression analysis, the model is no longer a simple linear one. That is known as multiple regression.

Formula for Simple Linear Regression

The two factors that are involved in simple linear regression analysis are designated x and y. The equation that describes how y is related to x is known as the regression model.

The simple linear regression model is represented by:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- β_0 is the y-intercept of the regression line.
- β_1 is the slope.
- E(y) is the mean or expected value of y for a given value of x.

The linear regression model contains an error term that is represented by ε. The error term is used to account for the variability in y that cannot be explained by the linear relationship between x and y. If ε were not present, that would mean that knowing x would provide enough information to determine the value of y.

NonLinear Regression

- While a linear equation has one basic form, nonlinear equations can take many different forms. The easiest way to determine whether an equation is nonlinear is to focus on the term “nonlinear” itself. Literally, it’s not linear. If the equation doesn’t meet the criteria above for a linear equation, it’s nonlinear.
- That covers many different forms, which is why nonlinear regression provides the most flexible curve-fitting functionality.

Equation for Linear and Nonlinear Regression

Linear regression uses a linear equation in one basic form, **$Y = a + bx$**, where x is the explanatory variable and Y is the dependent variable:

$$Y = a0 + b1X1.$$

If your equation looks like the examples above (i.e. it looks like $Y = a + bx$, it's linear. If not, it's nonlinear.

And you can even square a term to model a curve:

$$Y = a0 + b1X1\text{^}2.$$

A nonlinear regression equation can take on multiple forms. ⇒ **$Y = a + bx\text{^}2$**

Regression Equations

Linear Regression Formula:

Regression Equation(y) = a + bx

Slope(b) = $(N\Sigma XY - (\Sigma X)(\Sigma Y)) / (N\Sigma X^2 - (\Sigma X)^2)$

Intercept(a) = $(\Sigma Y - b(\Sigma X)) / N$

Non-linear Regression Formula:

Regression Equation(y) = $a + bx^2$

We can still use Linear Regression formula:

Slope(b) = $(N\Sigma \underline{P}Y - (\Sigma \underline{P})(\Sigma Y)) / (N\Sigma \underline{P}^2 - (\Sigma \underline{P})^2)$

Intercept(a) = $(\Sigma Y - b(\Sigma \underline{P})) / N$

Where $\underline{P}$ = X * X

Where:

- x and y are the variables.
- b = The slope of the regression line
- a = The intercept point of the regression line and the y axis.
- N = Number of values or elements
- X = First Score
- Y = Second Score
- ΣXY = Sum of the product of first and Second Scores
- ΣX = Sum of First Scores
- ΣY = Sum of Second Scores
- ΣX^2 = Sum of square First Scores

Linear Regression Equation - Steps

- To find Linear regression equation, we will first find slope, intercept and use it to form regression equation.

- Step 1:
  - ➢ Count the number of values. N = 10

- Step 2:
  - ➢ Find X * Y, X2

x	y	x*y	x*x
1	1.8	1.8	1
2	2.4	4.8	4
3.3	2.3	7.59	10.89
4.3	3.8	16.34	18.49
5.3	5.3	28.09	28.09
1.4	1.5	2.1	1.96
2.5	2.2	5.5	6.25
2.8	3.8	10.64	7.84
4.1	4	16.4	16.81
5.1	5.4	27.54	26.01

Linear Regression Equation - Steps Continues

- Step 3:

  Find $\Sigma \underline{X}$, ΣY, $\Sigma \underline{X}Y$, $\Sigma \underline{X}^2$

	x	y	x*y	x*x
	1	1.8	1.8	1
	2	2.4	4.8	4
	3.3	2.3	7.59	10.89
	4.3	3.8	16.34	18.49
	5.3	5.3	28.09	28.09
	1.4	1.5	2.1	1.96
	2.5	2.2	5.5	6.25
	2.8	3.8	10.64	7.84
	4.1	4	16.4	16.81
	5.1	5.4	27.54	26.01
Total	31.8	32.5	120.8	121.34

Linear Regression Equation - Steps Continues

- Step 4:
  Slope(b) = $(N\Sigma \underline{X}Y - (\Sigma \underline{X})(\Sigma Y)) / (N\Sigma \underline{X}^2 - (\Sigma \underline{X})^2)$
- Step 5: ===> Intercept(a) = $(\Sigma Y - b(\Sigma X)) / N$

Real Data Set 1 50% of the collcted data				Intercept(a) = (ΣY - b(ΣX)) / N	Slope(b) = (NΣXY - (ΣX)(ΣY)) / (NΣX2 -(ΣX)2)
x	y	x*y	x*x	a1	b1
1	1.8	1.8	1	0.505094797	0.863177681
2	2.4	4.8	4	0.505094797	0.863177681
3.3	2.3	7.59	10.89	0.505094797	0.863177681
4.3	3.8	16.34	18.49	0.505094797	0.863177681
5.3	5.3	28.09	28.09	0.505094797	0.863177681
1.4	1.5	2.1	1.96	0.505094797	0.863177681
2.5	2.2	5.5	6.25	0.505094797	0.863177681
2.8	3.8	10.64	7.84	0.505094797	0.863177681
4.1	4	16.4	16.81	0.505094797	0.863177681
5.1	5.4	27.54	26.01	0.505094797	0.863177681
31.8	32.5	120.8	121.34	0.505094974	0.863177681

Linear Regression Equation - Steps Continues

- Step 6:
  Then substitute Intercept(a) and Slope(b) in regression equation formula
  Regression Equation(y) = a + bx

Real Data Set 1 50% of the collcted data				Intercept(a) = (ΣY - b(ΣX)) / N	Slope(b) = (NΣXY - (ΣX)(ΣY)) / (NΣX2 -(ΣX)2)	Model 1: Linear Regression
x	y	x*y	x*x	a1	b1	ŷ=a1 + b1 * x
1	1.8	1.8	1	0.505094797	0.863177681	1.8
2	2.4	4.8	4	0.505094797	0.863177681	4.8
3.3	2.3	7.59	10.89	0.505094797	0.863177681	7.59
4.3	3.8	16.34	18.49	0.505094797	0.863177681	16.34
5.3	5.3	28.09	28.09	0.505094797	0.863177681	28.09
1.4	1.5	2.1	1.96	0.505094797	0.863177681	2.1
2.5	2.2	5.5	6.25	0.505094797	0.863177681	5.5
2.8	3.8	10.64	7.84	0.505094797	0.863177681	10.64
4.1	4	16.4	16.81	0.505094797	0.863177681	16.4
5.1	5.4	27.54	26.01	0.505094797	0.863177681	27.54
31.8	32.5	120.8	121.34	0.505094974	0.863177681	

Nonlinear Regression Equation - Steps

- To find Nonlinear regression equation, we will first find slope, intercept and use it to form regression equation.
- Step 1:

  Count the number

  of values. N = 10

- Step 2:

  Find X^2 * Y, X^2,

  X^4

Real Data Set 1 50% of the collcted data				
x	y	(x*x)*y	X^4	x*x
1	1.8	1.8	1	1
2	2.4	9.6	16	4
3.3	2.3	25.047	118.5921	10.89
4.3	3.8	70.262	341.8801	18.49
5.3	5.3	148.877	789.0481	28.09
1.4	1.5	2.94	3.8416	1.96
2.5	2.2	13.75	39.0625	6.25
2.8	3.8	29.792	61.4656	7.84
4.1	4	67.24	282.5761	16.81
5.1	5.4	140.454	676.5201	26.01

Nonlinear Regression Equation - Steps Continues

- Step 3:

  Find ΣX, ΣY, ΣX^2,ΣX^2Y,ΣX^4.

	Real Data Set 1 50% of the collcted data					Intercept(a) = (ΣY - b(ΣX)) / N
	x	y	x*x	(x*x)*y	X^4	a1
	1	1.8	1	1.8	1	0.505094797
	2	2.4	4	9.6	16	0.505094797
	3.3	2.3	10.89	25.047	118.5921	0.505094797
	4.3	3.8	18.49	70.262	341.8801	0.505094797
	5.3	5.3	28.09	148.877	789.0481	0.505094797
	1.4	1.5	1.96	2.94	3.8416	0.505094797
	2.5	2.2	6.25	13.75	39.0625	0.505094797
	2.8	3.8	7.84	29.792	61.4656	0.505094797
	4.1	4	16.81	67.24	282.5761	0.505094797
	5.1	5.4	26.01	140.454	676.5201	0.505094797
Total	31.8	32.5	121.34	509.762	2329.9862	0.505094974

Nonlinear Regression Equation - Steps Continues

- Step 4: Slope(b) = $(N\Sigma\underline{X}Y - (\Sigma\underline{X})(\Sigma Y)) / (N\Sigma\underline{X}^2 - (\Sigma\underline{X})^2)$
- Step 5: Intercept(a) = $(\Sigma Y - b(\Sigma\underline{X})) / N$

	Real Data Set 1 50% of the collcted data					Intercept(a) = (ΣY - b(ΣP)) / N ==> P = X * X ==> (ΣY - b(Σx*x)) / N	Slope(b) = (NΣPY - (ΣP)(ΣY)) / (NΣP2 - (ΣP)2) ==> P = X * X ==> (NΣx^2Y - (Σx^2)(ΣY)) / (NΣ(x*x)2 - (Σx*x)2)
	x	y	x*x	(x*x)*y	X^4	a2	b2
	1	1.8	1	1.8	1	1.617219705	0.134562411
	2	2.4	4	9.6	16	1.617219705	0.134562411
	3.3	2.3	10.89	25.047	118.5921	1.617219705	0.134562411
	4.3	3.8	18.49	70.262	341.8801	1.617219705	0.134562411
	5.3	5.3	28.09	148.877	789.0481	1.617219705	0.134562411
	1.4	1.5	1.96	2.94	3.8416	1.617219705	0.134562411
	2.5	2.2	6.25	13.75	39.0625	1.617219705	0.134562411
	2.8	3.8	7.84	29.792	61.4656	1.617219705	0.134562411
	4.1	4	16.81	67.24	282.5761	1.617219705	0.134562411
	5.1	5.4	26.01	140.454	676.5201	1.617219705	0.134562411
Total	31.8	32.5	121.34	509.762	2329.9862	1.617219705	0.134562411

Nonlinear Regression Equation - Steps Continues

Step 6:

Then substitute these values in regression equation formula

Regression Equation(y) = a + bx^2

Real Data Set 1 50% of the collcted data					Intercept(a) = (ΣY - b(ΣP)) / N ==> P = X * X ==> (ΣY - b(Σx*x)) / N	Slope(b) = (NΣPY - (ΣP)(ΣY)) / (NΣP2 - (ΣP)2) ==> P = X * X ==> (NΣx^2Y - (Σx^2)(ΣY)) / (NΣ(x*x)2 - (Σx*x)2)	Model 2: Non-Linear Regression
x	y	x*x	(x*x)*y	X^4	a2	b2	ŷ=a2 + b2 * x2
1	1.8	1	1.8	1	1.617219705	0.134562411	1.751782116
2	2.4	4	9.6	16	1.617219705	0.134562411	2.155469349
3.3	2.3	10.89	25.047	118.5921	1.617219705	0.134562411	3.082604361
4.3	3.8	18.49	70.262	341.8801	1.617219705	0.134562411	4.105278684
5.3	5.3	28.09	148.877	789.0481	1.617219705	0.134562411	5.39707783
1.4	1.5	1.96	2.94	3.8416	1.617219705	0.134562411	1.880962031
2.5	2.2	6.25	13.75	39.0625	1.617219705	0.134562411	2.458234774
2.8	3.8	7.84	29.792	61.4656	1.617219705	0.134562411	2.672189007
4.1	4	16.81	67.24	282.5761	1.617219705	0.134562411	3.879213834
5.1	5.4	26.01	140.454	676.5201	1.617219705	0.134562411	5.117188015
31.8	32.5	121.34	509.762	2329.9862	1.617219705	0.134562411	

Calculate y^ in Validation Phase

- Calculate y^ in Validation Phase is used to compare the performances of the prediction algorithms that were created based on the training set.

Validation Phase				
Real Data Set 2 25% of the collcted data			Model 1: Linear Regression	Model 2: Non-Linear Regression
x	y	x*x	ŷ=a1 + b1 * x ==> ŷ=0.5050947974 + 0.863177681 * x	ŷ=a2 + b2 * x2 ==> ŷ= 1.617219705+ 0.134562411 * x^2
1.5	1.7	2.25	1.799861319	1.91998513
2.9	2.7	8.41	3.0083100072	2.748889582
3.7	2.5	13.69	3.698852217	3.459379112
4.7	2.8	22.09	4.562029898	4.589703364
5.1	5.5	26.01	4.907300971	5.117188015
X	X	X	X	X
X	X	X	X	X
X	X	X	X	X
X	X	X	X	X
X	X	X	X	X

Time to Choose Best Model

- Choose the Model that has the best performance.

Training Phase				Validation Phase			
Real Data Set 1 50% of the collcted data		Model 1: Linear Regression	Model 2: Non-Linear Regression	Real Data Set 2 25% of the collcted data		Model 1: Linear Regression	Model 2: Non-Linear Regression
x	y	ŷ=a1 + b1 * x	ŷ=a2 + b2 * x2	x	y	ŷ=a1 + b1 * x	ŷ=a2 + b2 * x2
1	1.8	1.368272478	1.751782116	1.5	1.7	1.799861319	1.91998513
2	2.4	2.231450159	2.155469349	2.9	2.7	3.008310072	2.748889582
3.3	2.3	3.353581145	3.082604361	3.7	2.5	3.698852217	3.459379112
4.3	3.8	4.216758826	4.105278684	4.7	2.8	4.562029898	4.589703364
5.3	5.3	5.079936507	5.39707783	5.1	5.5	4.907300971	5.117188015
1.4	1.5	1.713543551	1.880962031	X	X	X	X
2.5	2.2	2.663039	2.458234774	X	X	X	X
2.8	3.8	2.921992304	2.672189007	X	X	X	X
4.1	4	4.04412329	3.879213834	X	X	X	X
5.1	5.4	4.907300971	5.117188015	X	X	X	X

Calculate MSE

- The Mean Squared Error (MSE) is a measure of how close a fitted line is to data points.
  - ➢ The smaller the MSE, the closer the fit is to the data.
- If Y is a vector of n predictions, and Y is the vector of the true values, then the (estimated) MSE of the predictor is:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2.$$

Training Set - MSE

Training Set -

Model 1 - MSE

= [(1.37-1.8)^2 + (2.155-2.4)^2 + (3.3535-2.3)^2 + (4.2167-3.8)^2

+(5.079-5.3)^2 +(1.714-1.5)^2 + (2.66-2.2)^2 + (2.9219-3.8)^2 +

(4.044-4)^2 + (4.9073-5.4)^2]/10 ⇒ 0.282254947

Model 2 - MSE

= [(1.75-1.8)^2 + (2.155-2.4)^2 + (3.083-2.3)^2 + (4.105-3.8)^2 +

(5.397-5.3)^2 +(1.8809-1.5)^2 + (2.458-2.2)^2 + (2.672-3.8)^2 +

(3.879-4)^2 + (5.117-5.4)^2]/10 ⇒ 0.235555579

TP - MSE M1		0.186388653
		0.028409049
		1.110033229
		0.173687919
		0.048427941
		0.045600848
		0.214405116
		0.770897514
		0.001946865
		0.242752333
	Sum =	0.28225494

TP - MSE M2	0.002324964
	0.059795239
	0.612469586
	0.093195075
	0.009424105
	0.145132069
	0.066685199
	1.271957636
	0.014589298
	0.079982619
Sum =	0.235555579

Validation Set - MSE

Validation Set -

Model 1 - MSE

= [(1.799-1.7)^2 + (3.008-2.7)^2 + (3.698-2.5)^2

+ (4.56-2.8)^2 + (4.907-5.5)^2]/5

= 0.999663104

Model 2 - MSE

= [(1.919-1.7)^2 + (2.748-2.7)^2 + (3.459-2.5)^2 +

(4.589-2.8)^2 + (5.117-5.5)^2]/5

= 0.864155015

VP - MSE M1	0.009972283
	0.0950551
	1.437246638
	3.104749361
	0.351292139
Sum =	0.999663104

VP - MSE M2	0.048393457
	0.002390191
	0.920408281
	3.203038131
	0.146545016
Sum =	0.864155015

Conclusion for Best Model

- ➢ The smaller the MSE, the closer the fit is to the data

max(Training_Set_MSE, Validation_Set_MSE) / min(Training_Set_MSE, Validation_Set_MSE)
Compare Model 1 and Model 2 -
Model 1 -- 0.999663104/0.282254947
3.541702691
Model 2 -- 0.864155015/0.235555579
3.668582247
Conclusion Model 1 is a better model

Test Set - Calculate Y With Best Model

- So, we apply our chosen prediction algorithm or model on our test set in order to see how it's going to perform so we can have an idea about our algorithm's performance on unseen data.

	Test Phase	
	Real Data Set 3 25% of the collcted data	Model is better ==> from Model 1 and Model 2 depending on the analysis of overfitting
	x	So, Linear Regression ŷ=a1 + b1 * x ==> Model 1 is best on the analysis of overfitting
3	1.4	1.713543551
2	2.5	2.663039
2	3.6	3.612534449
4	4.5	4.389394362
5	5.4	5.166254275
	X	X
	X	X
	X	X
	X	X
	X	X

Test Set with Best Model

- Finally,Y calculated in Test Phase with best model selected from Overfitting.

Training Phase				Validation Phase				Test Phase	
Real Data Set 1 50% of the collcted data		Model 1: Linear Regression	Model 2: Non-Linear Regression	Real Data Set 2 25% of the collcted data		Model 1: Linear Regression	Model 2: Non-Linear Regression	Real Data Set 3 25% of the collcted data	Model is better ==> from Model 1 and Model 2 depending on the analysis of overfitting
x	y	ŷ=a1 + b1 * x	ŷ=a2 + b2 * x2	x	y	ŷ=a1 + b1 * x	ŷ=a2 + b2 * x2	x	So, Linear Regression ŷ=a1 + b1 * x ==> Model 1 is best on the analysis of overfitting
1	1.8	1.368272478	1.751782116	1.5	1.7	1.799861319	1.91998513	1.4	1.713543551
2	2.4	2.231450159	2.155469349	2.9	2.7	3.008310072	2.748889582	2.5	2.663039
3.3	2.3	3.353581145	3.082604361	3.7	2.5	3.698852217	3.459379112	3.6	3.612534449
4.3	3.8	4.216758826	4.105278684	4.7	2.8	4.562029898	4.589703364	4.5	4.389394362
5.3	5.3	5.079936507	5.39707783	5.1	5.5	4.907300971	5.117188015	5.4	5.166254275
1.4	1.5	1.713543551	1.880962031	X	X	X	X	X	X
2.5	2.2	2.663039	2.458234774	X	X	X	X	X	X
2.8	3.8	2.921992304	2.672189007	X	X	X	X	X	X
4.1	4	4.04412329	3.879213834	X	X	X	X	X	X
5.1	5.4	4.907300971	5.117188015	X	X	X	X	X	X

References

- https://npu85.npu.edu/~henry/npu/classes/data_science/algorithm/slide/overfit.html
- https://npu85.npu.edu/~henry/npu/classes/data_science/algorithm/slide/linear_regression_example.html#lf
- https://npu85.npu.edu/~henry/npu/classes/data_science/algorithm/slide/non_linear_regression_example.html#nl
- Google Slides URL - https://docs.google.com/presentation/d/1YcYXMXg8SQbBT12w6R3MLeNMssGOGSk2SY6FlbKQ9TU/edit?usp=sharing
- GitHub URL -

  https://github.com/santhinagalla/Machine-Learning/tree/main/Model%20Selection/Use%20Overfitting%20To%20Evaluate%20Different%20Models